Conservation Biology

Registered Report

Mechanisms and impacts of an incentive-based conservation program with evidence from a randomized control trial

Emma Wiik,[1] Julia P. G. Jones,[1*] Edwin Pynegar,[1,2] Patrick Bottazzi,[1,4] Nigel Asquith,[3,5] James Gibbons,[1] and Andreas Kontoleon[6]

[1]School of Natural Sciences, Deiniol Road, Bangor University, Bangor, LL57 2UW, U.K.
[2]Fundacion Natura Bolivia, Calle Rio Totaitu 15, Santa Cruz de la Sierra, Bolivia
[3]Sustainability Science Program, Harvard Kennedy School, Cambridge, MA, 02138, U.S.A.
[4]Institute of Geography, University of Bern, Hallerstrasse 12, Bern, 3012, Switzerland
[5]Cuencas Sustentables, Calle Rio Totaitu 15, Santa Cruz de la Sierra, Bolivia
[6]Department of Land Economy, University of Cambridge, 19 Silber Street, Cambridge, CB3 9EP, U.K.

Abstract: Conservation science needs more high-quality impact evaluations, especially ones that explore mechanisms of success or failure. Randomized control trials (RCTs) provide particularly robust evidence of the effectiveness of interventions (although they have been criticized as reductionist and unable to provide insights into mechanisms), but there have been few such experiments investigating conservation at the landscape scale. We explored the impact of Watershared, an incentive-based conservation program in the Bolivian Andes, with one of the few RCTs of landscape-scale conservation in existence. There is strong interest in such incentive-based conservation approaches as some argue they can avoid negative social impacts sometimes associated with protected areas. We focused on social and environmental outcomes based on responses from a household survey in 129 communities randomly allocated to control or treatment (conducted both at the baseline in 2010 and repeated in 2015–2016). We controlled for incomplete program uptake by combining standard RCT analysis with matching methods and investigated mechanisms by exploring intermediate and ultimate outcomes according to the underlying theory of change. Previous analyses, focused on single biophysical outcomes, showed that over its first 5 years Watershared did not slow deforestation or improve water quality at the landscape scale. We found that Watershared influenced some outcomes measured using the survey, but the effects were complex, and some were unexpected. We thus demonstrated how RCTs can provide insights into the pathways of impact, as well as whether an intervention has impact. This paper, one of the first registered reports in conservation science, demonstrates how preregistration can help make complex research designs more transparent, avoid cherry picking, and reduce publication bias.

Keywords: causal inference, impact evaluation, matching, payments for ecosystem services, preanalysis plan, preregistration, robust, theory of change

Mecanismos e Impactos de un Programa de Conservación Basada en Incentivos con Evidencias de un Ensayo Aleatorio de Control

Resumen: Las ciencias de la conservación necesitan evaluaciones de impacto de mayor calidad, especialmente aquellas que exploran los mecanismos del éxito o del fracaso. Los ensayos aleatorios de control (RCTs) proporcionan evidencias particularmente sólidas de la efectividad de las intervenciones (aunque han recibido críticas por considerarlas reduccionistas e incapaces de proporcionar conocimiento sobre los mecanismos) pero ha habido pocos experimentos de ese tipo que investiguen los efectos de la conservación a escala del paisaje. Exploramos el impacto de Watershared, un programa de conservación basada en incentivos en marcha en los Andes bolivianos, con uno de los pocos RCTs existentes de conservación a escala de paisaje. Existe un gran interés por dichas

[*]*Address correspondence to J. P. G. Jones, email julia.jones@bangor.ac.uk*
***Article impact statement**: Randomized control trials can be used to evaluate program impacts and explore mechanisms through evaluation of intermediate outcomes.*

Paper submitted December 11, 2019; revised manuscript accepted February 7, 2020.

DOI: 10.1111/cobi.13508

estrategias de conservación basada en incentivos pues hay quienes argumentan que pueden evitar los impactos sociales negativos que a veces se asocian con las áreas protegidas. Nos enfocamos en los resultados sociales y ambientales con base en las respuestas de una encuesta a hogares en 129 comunidades asignadas al azar para controlar o tratar (ambas encuestas realizadas en la línea base en 2010 y repetidas en 2015/16). Impusimos un control para la aceptación incompleta del programa al combinar el análisis estandarizado de RCTs con métodos de emparejamiento e investigamos los mecanismos mediante la exploración de resultados intermedios y finales de acuerdo con la teoría subyacente del cambio. Los análisis previos, enfocados en resultados biofísicos únicos, mostraron que durante los primeros cinco años del programa Watershared, la deforestación no experimentó una desaceleración y tampoco hubo mejoras en la calidad del agua a escala de paisaje. Descubrimos que Watershared influyó sobre algunos resultados medidos con la encuesta, pero sus efectos fueron complejos y algunos fueron inesperados. De este modo demostramos como los RCTs pueden proporcionar conocimiento sobre las vías de impacto, así como también si una intervención genera un impacto. Este artículo, uno de los primeros reportes registrados en las ciencias de la conservación, demuestra cómo el prerregistro puede ayudar a hacer más transparentes los diseños complejos de investigación, evitar la selección subjetiva de datos y reducir el sesgo de publicación.

Palabras Clave: emparejamiento, evaluación de impacto, inferencia causal, pagos por servicios ambientales, plan de pre-análisis, prerregistro, sólido, teoría de cambio

摘要: 保护科学需要更多高质量的影响评估, 特别是那些探索成功或失败机制的评估。随机对照试验为探讨干预措施的有效性提供了稳健证据 (尽管被批评为简化方法且无法提供对机制的深入了解), 但在景观尺度上研究保护问题的此类试验还很少。我们探索了玻利维亚安第斯山脉基于激励措施的一项保护项目—Watershared的影响, 它是目前少有的景观尺度保护的随机对照试验之一。人们对这种基于激励机制的保护方法很感兴趣, 因为一些人认为它们可以避免与保护区有关的某些负面社会影响。我们在129个社区通过随机分配对照组和实验组进行了家庭调查 (包括2010年基线调查和2015/16年重复调查), 重点关注该项目的社会和环境结果。我们通过结合标准随机对照试验分析与匹配方法来控制不完整的调查, 并根据变化理论来通过分析中间结果和最终结果确定了相应机制。之前的分析集中在单一生物物理的结果, 发现在最初的5年内Watershared项目并没有在景观水平上减缓森林砍伐或改善水质。而我们发现该项目影响了本调查关注的一些结果, 但影响是复杂的, 且有些影响超出预料。因此, 我们展示了随机对照试验如何帮助深入了解影响的途径及干预措施是否产生了影响。本论文作为保护科学中最早的登记报告之一, 展示了预登记如何有助于使复杂的研究设计透明化, 避免人为选取结果, 并减少论文发表的偏倚性。**【翻译: 胡怡思; 审校: 聂永刚】**

关键词: 预分析计划, 预登记, 生态系统服务付费, 因果推理, 影响评估, 匹配, 变化理论, 稳健

Introduction

There is considerable interest in using positive incentives to encourage sustainable land management, conserve forests, and protect biodiversity. Those promoting these incentive-based conservation approaches, which include payments for ecosystem services (PES) (Jack et al. 2008), suggest they can both effectively deliver environmental outcomes and result in better social outcomes than strict protected areas (Sims & Alix-Garcia 2017). Synthesis of the existing evidence base suggests PES-type interventions have, if anything, only a modest impact on environmental outcomes, and impacts on social outcomes are even more uncertain (Samii et al. 2014; Liu & Kontoleon 2018). More and better quality evaluations are needed, especially those that can cast light on the mechanisms by which outcomes are, or are not, delivered (Miteva et al. 2012; Börner et al. 2016, 2017).

Randomized control trials (RCT) randomly allocate experimental units to treatment and control groups and are therefore often considered to provide particularly robust evidence of the effectiveness of interventions (Ferraro 2009). However, in the context of conservation policies, RCTs are rare (Pynegar et al. 2019). To our knowledge there have been 2 RCTs of incentive-based conservation interventions implemented at scale. Jayachandran et al. (2017) showed that carbon payments slowed deforestation rates in Uganda. The RCT in Bolivia of the Watershared intervention (Bottazzi et al. 2018) has been used to evaluate the impact of incentivizing farmers to keep cattle out of riparian forest and reduce deforestation on water quality (Pynegar et al. 2018), deforestation rates (Wiik et al. 2019), and environmental values (Grillos et al. 2019). A third landscape-scale RCT in conservation explored the impact of unconditional livelihood payments on deforestation rates in Sierra Leone (Wilebore et al. 2019).

Evaluation of such socioecological interventions is inherently complex because whether or not the incentives and associated social processes will produce the desired land-use change is uncertain and, even if achieved, these land-use changes may (or may not) result in the desired social and environmental ultimate outcome. Impacts may also differ between strata of society (Daw et al. 2016) and take time to materialize. There is interest in other disciplines, such as public health, in bringing lessons from qualitative impact evaluation into RCT analyses (Bonell et al. 2012). In qualitative impact evaluation, the focus is on building and validating a theory of change (which identifies the mechanisms by which the intervention delivers intermediate and ultimate outcomes of interest [White 2009]) rather than a narrow focus on ultimate outcomes. Published studies in which an RCT was used to evaluate the impact of landscape-scale conservation interventions mostly report ultimate environmental outcomes of the intervention (e.g., deforestation rates) but say little about social outcomes (and how these might differ among groups) and the causal linkages between the intervention and intermediate and ultimate outcomes.

The Bolivian organization Natura Bolivia began to develop the incentive-based conservation program now known as Watershared in 2003 (Asquith & Vargas 2007). Watershared aims to establish a reciprocal relationship between environmental service users (municipal governments and water cooperatives) and services providers (upstream farmers and cattle owners) by using in-kind incentives for forest protection and exclusion of cattle from riparian forest to protect biodiversity and improve downstream water quality (Bottazzi et al. 2018). As of 2016, Watershared had 210,000 ha (4500 households) under conservation agreements (Asquith 2016). The Watershared RCT presented a rare opportunity to fully analyze the impacts of an incentive-based conservation program. Using a large household survey conducted at baseline (in 2010) and end line (in 2015–2016), we explored both intermediate outcomes (e.g., perceived importance of forest, livelihood changes, and cattle exclusion from riparian forest) and indicators of ultimate outcomes (e.g., perceived forest condition, incidence and frequency of diarrhea). We used the theory of change underpinning the intervention to structure the evaluation. This paper is submitted as a registered report (Parker et al. 2019).

Methods

Watershared RCT

In 2010 Natura applied Watershared in a new protected area (Area Natural de Manejo Integrado Río Grande y Valles Cruceños) as an RCT (Fig. 1) to evaluate the impact of the intervention on deforestation rates, quality and quantity of water available for local communities, environmental values, and local livelihoods. They selected the 129 communities in the 5 main municipalities overlapping the protected area, and these were randomly allocated to control (conservation agreements not offered) or treatment (agreements offered) subsequent to blocking by municipality, community size, and cattle numbers. Consent to conduct the trial was granted by municipal mayors on the understanding that the program would subsequently be implemented in all communities. The experiment was not blinded because participants unavoidably knew whether they belonged to a control or treatment community. In 2016 the experiment ended, and agreements were offered in control communities.

The Watershared intervention operates through combining incentives with environmental education; a key feature of the intervention is promoting the message that watershed protection is in everyone's interest (Bottazzi et al 2018). Natura gave an environmental education presentation to all treatment and control communities prior to recruiting treatment participants, so the randomization primarily tested the effect of the incentives. Reinforcement of the education messages would have occurred more strongly in treatment communities, where there were multiple visits to offer and monitor the conservation agreements from 2011 to 2015.

Watershared Agreements

There were 3 levels of Watershared agreements. Level-1 and level-2 agreements applied to forested land within 100 m of a stream or waterway, and level-3 agreements applied to any forested land (details in Bottazzi et al. [2018]). In all 3 levels, land clearance or timber extraction was not permitted. In addition, cattle had to be excluded from land under level-1 agreements, whereas level 2 required working toward removing cattle. The value of incentive packages ranged from the equivalent of US$1/ha/year to US$10/ha/year, and farmers with level-1 agreements received an additional US$100 worth of in-kind incentives at signing. Transportation costs of the materials to communities were covered by the program. Agreements were for an initial 3 years, were renewable, and were offered in treatment communities twice per year. Program technicians monitored level 1 and level 2 land annually by walking transects across the parcels to verify compliance; level 3 agreements were monitored using remote sensing (forest cover). Where blatant noncompliance was detected, the materials farmers had been given were removed and redistributed to the community. As with many incentive-based conservation interventions, not all conservation funded with Watershared agreements is additional (i.e., some would have happened anyway in the absence of the scheme, a common issue with PES-type programs [Ezzine-de-Blas et al. 2016]). Bottazzi et al. (2018) estimated that a maximum of 30% of agreements to exclude cattle and 14% to avoid deforestation appear to be additional.

Watershared Household Survey

The household survey was a structured questionnaire with >100 questions (Bottazzi et al. 2017). The

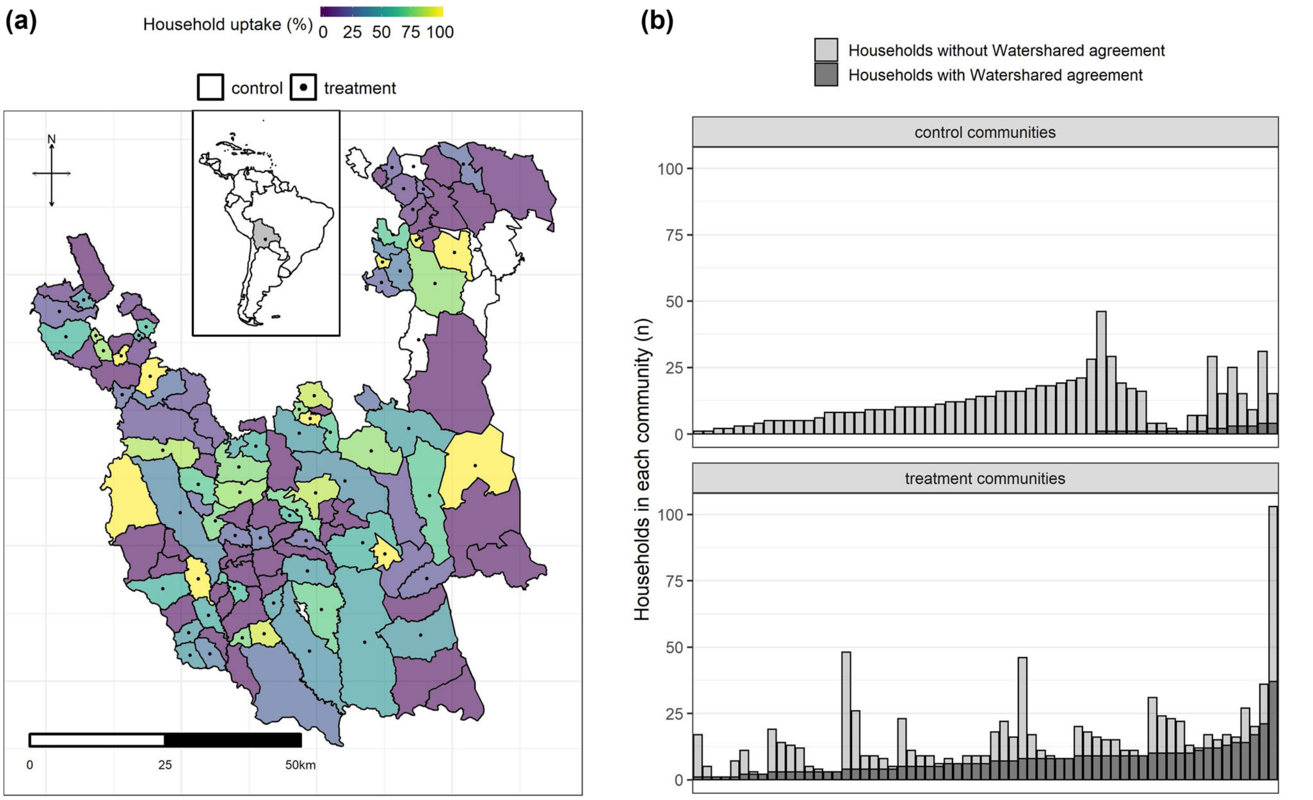

Figure 1. (a) Locations of the 64 control communities (Watershared agreements not offered) and 65 treatment communities (Watershared agreements offered) within the Area Natural de Manejo Integrado Río Grande y Valles Cruceños protected area (white, communities for which there are no households with both baseline and end line data [omitted from analyses]). (b) Distribution of the number of households per community and the number that had Watershared agreements in control (top) and treatment (bottom) communities, ordered by number of households with agreements.

baseline household survey was carried out by Natura in 2010 and the end line survey by Natura and Bangor University staff from October 2015 to June 2016. The aim was to deliver the baseline household survey to all households in all communities (Bottazzi et al. 2017). This was mostly achieved; the baseline reached 2623 households and only 57 previously unsurveyed households were found in the end line survey (Supporting Information). However, the end line survey was incomplete (Supporting Information) because 8 communities did not have any households with data from both baseline and end line surveys; these were excluded from the analyses (Fig. 1a).

Out of all households surveyed in both baseline and end line in treatment communities ($n = 970$), 456 households took up *Watershared* agreements and 514 did not (i.e., 47% uptake). The allocation to control and treatment was not perfect as 32 out of 702 households in control communities had agreements (Fig. 1b); however, 28 of these agreements were in land owned in treatment communities. Uptake percentages varied across communities from 0% to 100%, which in part reflects high percent uptake in a few small communities (Fig. 1b). Uptake of the program was influenced by barriers to entry (Grillos 2017), individual motivations (Bottazzi et al. 2018), and whether or not households were available to attend the meetings in which the program was presented (Wiik et al. 2019).

The consent form used in both baseline and end line surveys is archived alongside the data (Bottazzi et al. 2017). The end line survey was assessed under the Bangor University Research Ethics Framework. Natura were involved in the research (and paid the enumerators), which is a potential conflict of interest because they are also the implementers of the Watershared program that this work evaluates. However, the independent Bangor University team trained the enumerators, designed the survey, managed and cleaned the data, and conducted the analyses.

Selection of Outcome Variables

There are a large number of potential outcome variables from the survey that could be explored. We

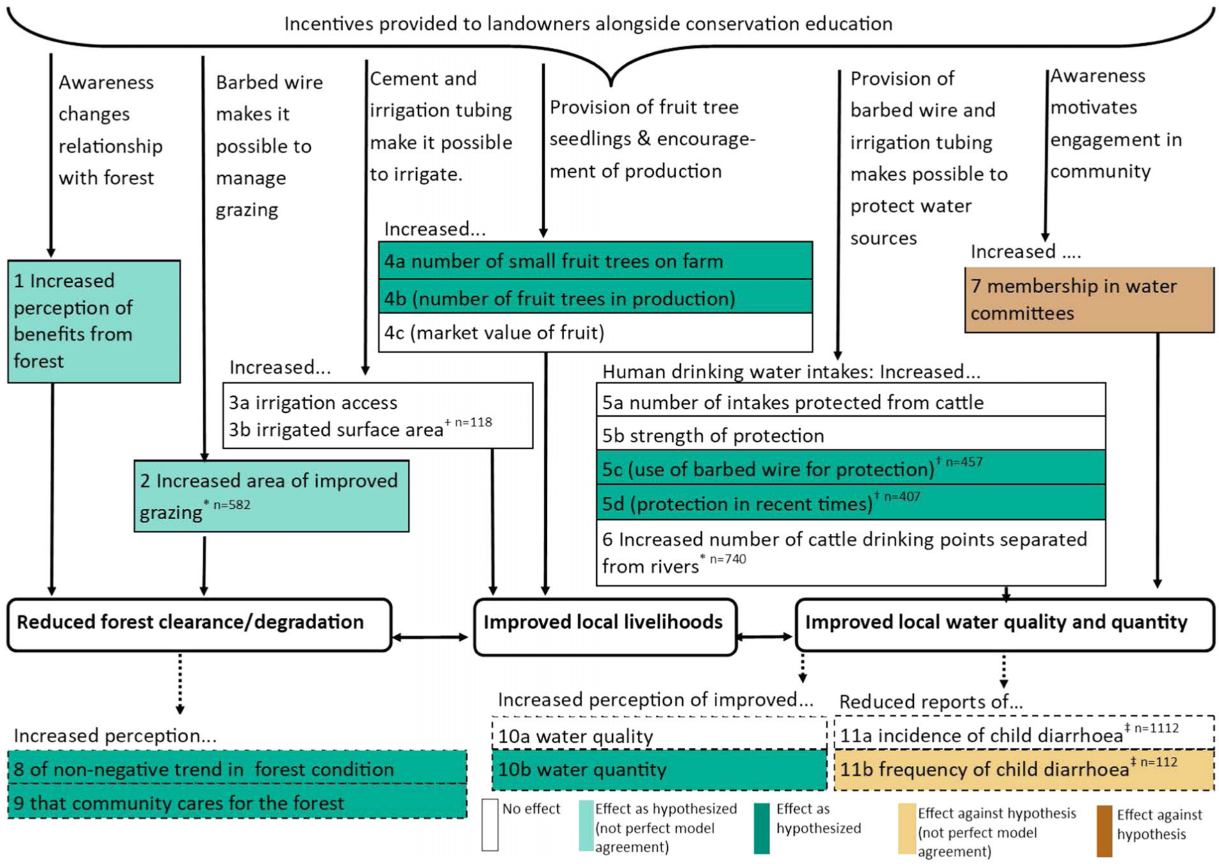
Incentives provided to landowners alongside conservation education
Awareness changes relationship with forest
Barbed wire makes it possible to manage grazing
Cement and irrigation tubing make it possible to irrigate.
Provision of fruit tree seedlings & encouragement of production
Provision of barbed wire and irrigation tubing makes possible to protect water sources
Awareness motivates engagement in community
1 Increased perception of benefits from forest
Increased...
4a number of small fruit trees on farm
4b (number of fruit trees in production)
4c (market value of fruit)
Increased ....
7 membership in water committees
Increased...
3a irrigation access
3b irrigated surface area+ n=118
2 Increased area of improved grazing* n=582
Human drinking water intakes: Increased...
5a number of intakes protected from cattle
5b strength of protection
5c (use of barbed wire for protection)† n=457
5d (protection in recent times)† n=407
6 Increased number of cattle drinking points separated from rivers* n=740
Reduced forest clearance/degradation
Improved local livelihoods
Improved local water quality and quantity
Increased perception...
8 of non-negative trend in forest condition
9 that community cares for the forest
Increased perception of improved...
10a water quality
10b water quantity
Reduced reports of...
11a incidence of child diarrhoea‡ n=1112
11b frequency of child diarrhoea‡ n=112
No effect
Effect as hypothesized (not perfect model agreement)
Effect as hypothesized
Effect against hypothesis (not perfect model agreement)
Effect against hypothesis

Figure 2. The simplified theory of change linking the Watershared intervention with intermediate outcomes (square boxes), ultimate outcomes (rounded boxes), and indicators of these ultimate outcomes (square boxes with dashed lines). The hypothesized direction of the effect of the intervention is indicated for each outcome tested. Some analyses are only relevant for a subset of data (∗, households who own cattle; +, households who have irrigation access; †, households reporting protected drinking water intakes; ‡, households with children and personal water intake; brackets, outcomes for which we expect limited impact [e.g., number of fruit trees in production may not yet be affected because they will not yet have had time to mature]). Colors show results of the regression analyses for as-treated models only (Supporting Information shows the same results with the as-randomized models) (green, results that support hypothesis; browns, results against the hypotheses). Less saturated colors show outcomes for where there was some disagreement in the significance between the models used as robustness checks (Fig. 4).

systematically selected outcome variables for analysis based on there being a clear hypothesized mechanism linking to Watershared objectives (Supporting Information), based on the program's underlying theory of change (Fig. 2). The outcome variables selected included intermediate outcomes that contribute to the attainment of Watershared ultimate outcomes (e.g., number of water intakes protected from cattle, perception that forest delivers benefits) and self-reported indicators of the ultimate Watershared outcomes (e.g., diarrheal disease in children, perception of forest condition). In total, we identified 11 main outcome variables of interest (some of which have more than 1 indicator) (Fig. 2).

One outcome had been evaluated previously. Pynegar (2018) found no effect of the intervention as implemented on diarrheal incidence and frequency. We accept this finding and do not reanalyze. However, we conducted a secondary analysis exploring the impact of the intervention on the subset of households that reported having an individual water intake, which households plausibly have more control over.

Four outcomes were analyzed for a subset of households (Fig. 2) rather than the full data set. Diarrhea frequency and incidence among children was only analyzed for households that have their own drinking water intake and have children. Hectares of irrigated land were only analyzed for those who reported having access to irrigation in both baseline and end line surveys. The extent and method by which water intakes are protected was only analyzed for those who reported protecting their water intake in the end line survey (there was not a baseline question on this variable). Hectares

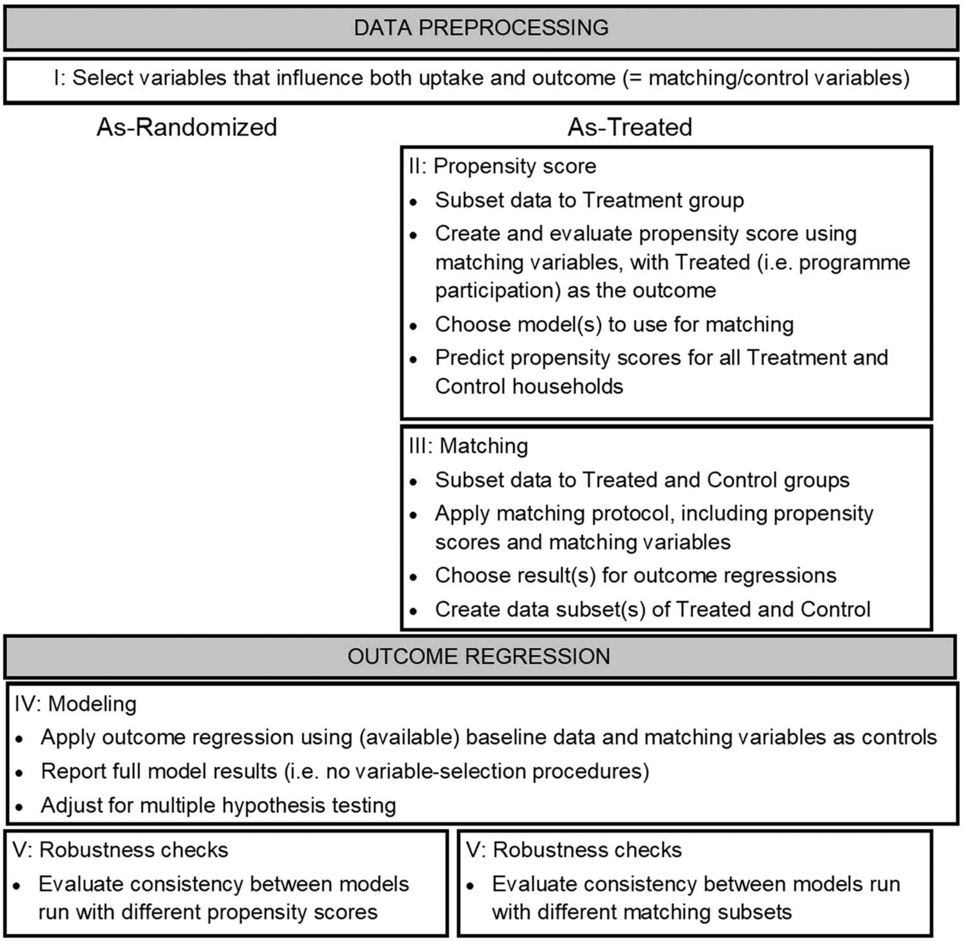
DATA PREPROCESSING
I: Select variables that influence both uptake and outcome (= matching/control variables)
As-Randomized
As-Treated
II: Propensity score
• Subset data to Treatment group
• Create and evaluate propensity score using matching variables, with Treated (i.e. programme participation) as the outcome
• Choose model(s) to use for matching
• Predict propensity scores for all Treatment and Control households
III: Matching
• Subset data to Treated and Control groups
• Apply matching protocol, including propensity scores and matching variables
• Choose result(s) for outcome regressions
• Create data subset(s) of Treated and Control
OUTCOME REGRESSION
IV: Modeling
• Apply outcome regression using (available) baseline data and matching variables as controls
• Report full model results (i.e. no variable-selection procedures)
• Adjust for multiple hypothesis testing
V: Robustness checks
• Evaluate consistency between models run with different propensity scores
V: Robustness checks
• Evaluate consistency between models run with different matching subsets

Figure 3. Outline of methods workflow for the as-randomized models and as-treated models. Phases I, II, and III show preprocessing undertaken for the stage-1 registered report (they were completed before initial peer review). Phases IV and V occurred at stage 2 of the registered report.

of improved grazing land were analyzed only for those who said they owned cattle in baseline and end line surveys.

Data Analyses

Our data analyses focused on testing, within the theory-of-change framework, the individual hypotheses within the 11 outcome categories of survey questions (Fig. 2). Within each category, we identified 1 main analysis where we would expect to see a change driven by the intervention if successful and in some cases also secondary analyses where we would expect it to be too early to detect change or where the change was indicative more of a detail in a process than an overarching mechanism or success of Watershared (Fig. 2). In our analyses, we followed a hierarchy by which the main analysis within a category was given most weight in evaluating the program (e.g., whether intakes were protected more or less in treatment communities was more important than when intakes were protected). The results from all analyses were evaluated against the theory-of-change logic. Where results conflicted with this logic, we evaluated the strength of evidence based on robustness checks. If results were robust, this cast doubt on the theory of change.

We tested our hypotheses using 2 analytical approaches, 1 estimated the average treatment effect (Glennerster & Takavarasha 2013) of the program as it was rolled out and the other estimated the program effect specifically on those who participated (Fig. 3). The first, as-randomized analysis, compared outcomes in all households in treatment communities with all households in control communities, regardless of uptake. The second, as-treated analysis, compared outcomes in treated households (households in treatment communities who participated, regardless of which incentives they selected [Supporting Information]) with statistically matched control households (matched control = households in control communities likely to have participated had they had the opportunity, excluding those who signed agreements). The distinction between the as-randomized and as-treated analyses is important due to the incomplete uptake of Watershared. For example, overall impacts of interventions may be low not because the intervention lacks efficacy but because of low levels of uptake or poor implementation and compliance (Glennerster & Takavarasha 2013).

Table 1. Variables selected for matching variables and control variables in the final-outcome models, indicating which were interacted with the experimental group (treatment or control, or treated or matched control) in outcome regressions (stage 2).

Variable	*Category*	*Mechanism*	*Outcomes for which variable is interacted with experimental group*
Community work frequency (n/year)	community cohesion	likely to be related to motivation to participate and adhere to agreements due to social norms	NA
Generations in a community (n)	community cohesion	likely to be related to level of engagement in the community and also ability to participate and follow through with agreements	NA
Land owned (ha)	wealth	likely to be related to ability to afford to invest time and effort in conservation	water committee membership; Diarrhea; Irrigation implementation
Forest ownership (binary)	wealth	likely to be related to owning eligible land and being able to afford to invest time and effort in conservation	NA
Cattle owned (n)	wealth	likely to be related to ability to afford to invest time and effort in conservation	cattle and human drinking water management; Improved grazing
Number of rooms in home	wealth	likely to be related to ability to afford to invest time and effort in conservation	NA
Education level (approx. years)	education	likely to be related to capacity to engage with the conservation program	cattle and human drinking water management; improved grazing; diarrhea; irrigation implementation; water committee membership
Perceived benefits from forest (binary)	environmental attitudes	related to motivation to engage with conservation	NA
Perceived problems in water quality, quantity (binary)	environmental attitudes	related to motivation to engage with conservation	human and cattle drinking water management

Abbreviation: NA, variables not interacted.

Some of our perception-based variables represent an observation of community-wide change and so blur the distinction between as-randomized and as-treated analyses (e.g., a perception of how the community is managing its forest can be the same for a participating and nonparticipating household). In these cases, the difference between the as-randomized and as-treated analyses tests whether participation in the program changes how a person perceives their environment.

Before stage 1 review of this registered report, we completed 3 phases of data preprocessing (Fig. 3). Phase I involved choosing variables for use in matching (in the As-Treated analyses) and for use as control variables in the final outcome regressions (both analyses). We selected variables that we hypothesized to influence both uptake and outcomes of the program. Candidate variables were considered based on previous work exploring the uptake of the Watershared intervention (Grillos 2017; Bottazzi et al. 2018; Wiik et al. 2019). We avoided variables with a lot of missing data (Supporting Information). The final set included variables capturing community cohesion, wealth, education, and predisposing environmental attitudes (Table 1). Baseline data for an outcome, where available, were used as control variables in outcome regressions as per some difference-in-differences analyses, but not as matching variables to avoid regression to the mean (Daw & Hatfield 2018) (S4). In phase II (Supporting Information), we developed propensity score models, based on the variables selected, to predict selection bias for households in control communities based on modeled participation in the program among households in treatment communities. In phase III (S6), we used the selected variables and the propensity scores (a primary and secondary version) to match treated households with the best available counterfactuals from the control households through a genetic matching algorithm. We used the R packages Matching (Sekhon 2011) to perform the matching, cobalt (Greifer 2020) to evaluate balance visually, mgcv for regressions (Wood 2011, 2017), and ggplot2 for plots (Wickham 2016).

The final 2 phases (outcome regressions, phase IV, and robustness checks, phase V) were carried out after stage 1 review was complete. Because we tested many outcomes, there was an increased probability of encountering at least 1 false positive (finding a significant impact on an outcome when there is, in fact, none). We therefore applied the Benjamini and Hochberg method to control the false detection rate at a level of 0.05. We ranked p-values based on the p-value from the primary analysis within outcome categories (Fig. 2) (see Supporting Information for full description of the multiple testing procedures). These methods were reviewed as part of our stage 1 plan.

The regressions used for hypothesis testing (as opposed to robustness checks) were those that included the primary propensity score (Supporting Information) and, in the case of matched analyses, the regressions run on the least restrictive caliper while still attaining adequate balance (a caliper limits the difference between any 1 pair of observations to within a given SD, meaning treated observations deemed too different from any 1 control were discarded) (Supporting Information). Regressions including the secondary propensity score and additional matching outputs were used as robustness checks (Supporting Information) (Ho et al. 2007). For example, we would not expect robust results to change if we used a slightly different set of control observations or a subset of treated households (where a caliper results in losing treated households).

In the as-randomized analyses (phase IV), the outcome was regressed on the experimental group (control or treatment) plus control variables, including the baseline data for an outcome where available. Our control variables included those used in matching to control for nonindependence of observations (Wan 2019), add precision to our effect estimate, correct for remaining biases (Ho et al. 2007; Hill 2008; Streiner 2015), and allow evaluation of heterogeneous treatment effects based on variable interactions (Ferraro & Hanauer 2014). All regressions were undertaken using generalized additive models (GAMs) (Wood 2017); families were fitted to the response expectation (e.g., binomial family with a logit link for binary outcomes [Supporting Information]).

As-treated regressions were similar except for being undertaken on the matched control and Treated subsets of data as per the matching protocol. The protocol resulted in 4 possible data sets: matching with and without a caliper and matching with 2 versions of the propensity score (Supporting Information). For the outcomes that were analyzed with the full data set (Fig. 2), we tested all 4 data sets in 4 regressions. For the outcomes only appropriate to explore with a smaller subset of data (e.g., those who own cattle or have children [Fig. 2]), we ran only 2 regressions because applying a caliper resulted in losing too many treated observations (S6).

To explore the extent to which the intervention may benefit socioeconomic groups differently and our expectation that some outcomes may be more feasible to achieve for some households than others, we explored a number of outcome interactions based on education and wealth indicators (Table 1). We also included an interaction between perception of water quantity or quality and the experimental group (control or treatment or matched control or treated) to examine whether the program had different impacts on those who are more influenced by these issues (Table 1).

Deviations from Preregistration

We undertook all outcome analyses with truncated values of highly skewed predictors, contrary to what was stated in Supporting Information of the stage-1 registered report. This was because we believed this added an unnecessary complication (testing whether outliers were biasing our estimates for each of 98 individual models).

Results

Checks suggested results were robust because there were no inconsistencies in the direction of effects for any models (Supporting Information). Robustness checks also confirmed the significance (or lack of) of the main analysis for as-randomized analyses. There was slightly less agreement in the significance for as-treated results (Fig. 4). This may have been because power was reduced in as-treated results due to lower sample sizes.

When presenting results, we talk both about treated households and treatment households. Treated households were those in treatment communities that signed Watershared agreement. They were always compared against a counterfactual of households in control communities matched on socioeconomic predictors of uptake of Watershared agreements. These results are those from the as-treated analysis. Treatment households were all those in treatment communities. They were always compared against households in control communities (without matching). These results derive from the as-randomized analysis.

For some outcomes there were significant treatment effects in the direction hypothesized (Figs. 2 & 5; Supporting Information). Treated households and treatment households had significantly more small fruit trees (mean of 50 and 25, respectively) and more fruit trees in production (mean of 100 and 150, respectively) than their counterfactuals. Treated and treatment households were also more likely to perceive positive trends over the last 5 years in water quantity and forest condition and more likely to perceive that the wider community cared more about the forest (Figs. 2 & 5; Supporting Information). The intervention may also have increased the area of improved grazing land; the results of the main model were not significant following *p*-value correction, but models used in the robustness checks showed a significant effect (Figs. 2 & 4; Supporting Information).

We found no evidence of a treatment effect on whether a household perceived gaining benefits from forest (Figs. 2 & 5; Supporting Information). However, given that the vast majority (over 90% in all groups) of respondents perceived benefits at baseline, there was little scope for increase. Nor were there treatment effects on irrigation access or irrigated land extent or perceptions of changes in water quality over time.

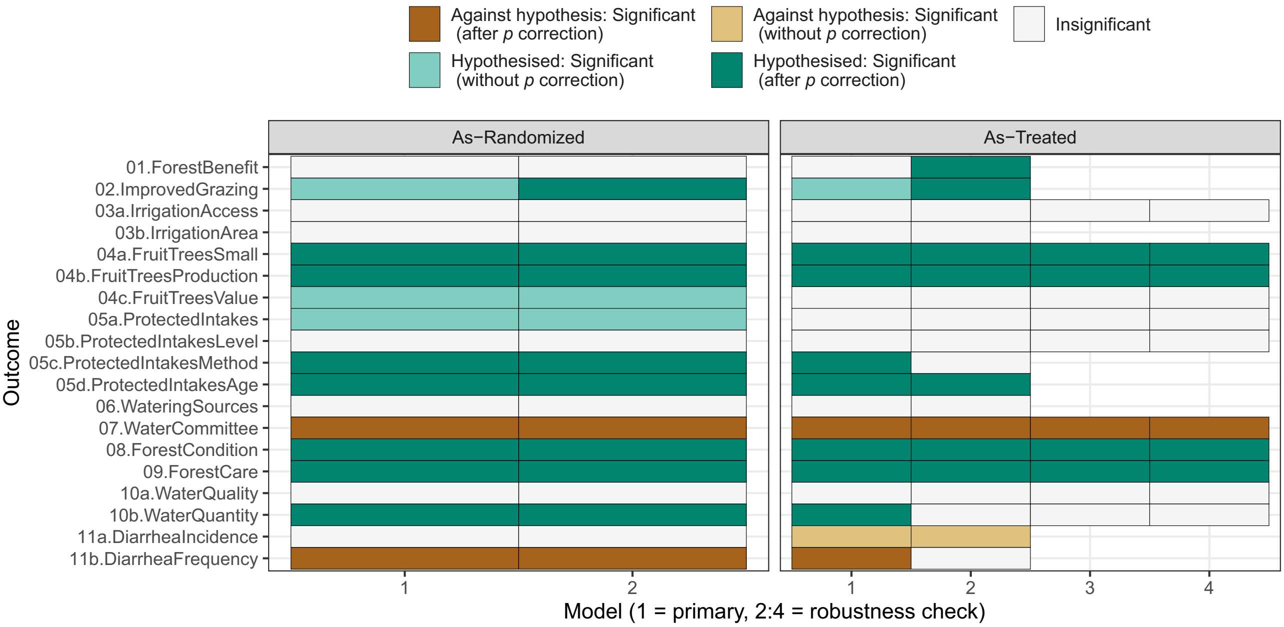

Figure 4. Robustness checks based on comparing the significance of intercept differences between control and treatment groups for the as-randomized and as-treated analyses for the primary model and subsidiary models (max 4 in as-treated due to matching protocols). The p-value correction was made based on the Benjamini–Hochberg threshold (see Methods).

There was no convincing treatment effect (i.e., effects were not significant after *p*-value correction) related to whether intakes were protected from cattle or to strength of protection of main water intakes from cattle access (Figs. 2 & 4). However, treated and treatment households were more likely to use barbed wire than traditional methods to protect intakes and to have implemented protected intakes more recently than their counterfactuals.

For some outcomes, there were significant treatment effects in the opposite direction to our hypotheses. At end line, control respondents were about 20% more likely to be members of water committees than treated or treatment households (Figs. 2 & 5). There was weaker evidence of a treatment effect against hypotheses for outcomes associated with diarrhea. The frequency of diarrhea was higher in treatment groups in both analyses, although the effect was not significant in robustness checks in the as-treated results (Figs. 2 & 4). There was also some evidence that incidence of diarrhea was higher in the treatment group, although this was not significant after *p*-value correction, and the effect was not seen in the as-treated results. This result may be an artefact of subsample bias or a lack of power in this subgroup. The diarrhea analysis was conducted on a small subset of the data (only households with children and their own water intake). In the as-randomized results, only 7 incidents of diarrhea were reported in the control group ($n = 61$); this was further reduced after matching in the as-treated analyses. It followed that in some models there was perfect separation.

Discussion

Data analyses involve multiple decisions as researchers seek to reveal the truth from complex, often messy, data (Fraser et al. 2018). Studies revealing a lack of reproducibility in fields such as preclinical medicine (Freedman et al. 2015) and psychology (Open Science Collaboration 2015) have resulted in much needed scrutiny of how these decisions are vulnerable to confusion, or even corruption. Although conservation science has so far avoided a scandal of reproducibility, a recent study of researchers in ecology and evolution (Fraser et al. 2018) revealed a worrying prevalence of cherry picking (failing to report results that are not significant) and reporting unexpected findings as if they were hypothesized from the start (hypothesizing after results are known [HARKing]). Preregistration of analysis can avoid these problems as long as the study is adequately powered to detect differences of interest. Submitting planned research for peer review as a registered report goes 1 step further and also reduces publication bias (Parker et al. 2019). The multiple outcomes available for analysis from the Watershared RCT were inevitably vulnerable to cherry picking and HARKing. By publishing this as a registered report, we reduced both the temptation to use, and the impression we may have used, questionable

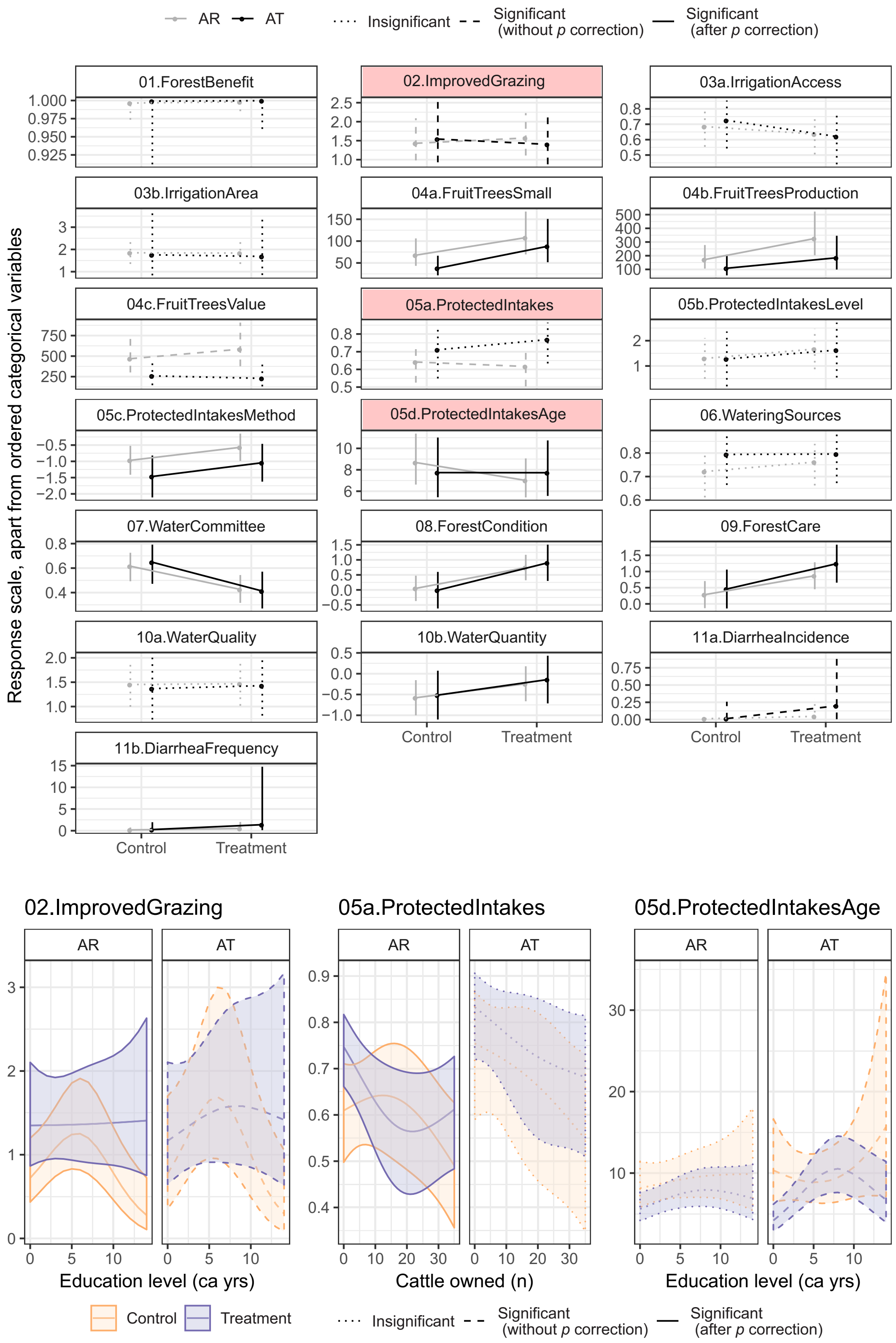

Figure 5. The differences in intercept values (line graphs with confidence bars) and interaction slopes (line graphs with confidence bands, where the x-axis is a continuous variable) between control and treatment groups for both as-randomized (AR) and as-treated (AT) analyses. Only interactions where at least one analysis shows significance are shown. Predictions are based on mean values for continuous predictors and common values for factor predictors (e.g., perceiving benefits from forest). The 95% CIS relate to the entire prediction, rather than the control–treatment difference.

research practices to tell a better story. Ideally, of course, preregistration should precede data collection. Data collection for this RCT began in 2010 and was complete in 2015, before preregistration was widely advocated. However, we submitted stage 1 before looking at outcome variables, meaning the study was accepted in principle based on the introduction and methods alone. This study is one of the first registered reports in conservation science.

Although large-scale RCTs of interventions are receiving increasing attention (e.g., the 2019 economics Nobel prize was awarded to Kremer, Banerjee, and Duflo for their experimental work on alleviating poverty), they remain rare in conservation (Pynegar et al. 2019). As far as we know, we are the first to use a RCT to look at outcomes from across the theory of change to give insights into the mechanism by which a conservation intervention works, or does not work. Watershared's ultimate aims are to reduce the rate of forest clearance and degradation, improve livelihoods, and improve water quality and quantity. Our previous analyses of the intervention, looking simply at biophysical measures of ultimate outcomes, revealed minimal impact on deforestation (Wiik et al. 2019) and water quality (Pynegar et al. 2018). Those analyses alone said little about why the intervention may not have resulted in a change in those outcomes or whether measurable impact might be detected given time. Looking closely at intermediate outcomes, as we did here, provides valuable insights to answer such questions about mechanisms.

Watershared aims to conserve forest by increasing the awareness of the benefits forests provide and increasing farmers' investment in improved grazing (reducing forest grazing) and alternatives to cattle ranching (e.g., fruit production). Over 90% of respondents already perceived benefits from forests, so it is unsurprising that the intervention did not increase this. The intervention appears to have increased the area of improved grazing and significantly increased fruit tree production (although this was not yet apparent in market values, which we predicted owing to lags in fruit tree maturation). Watershared also provided cement and irrigation tubing to increase irrigated agriculture. However, due to their relatively high cost, they were less popular than barbed wire and fruit trees (and field observations suggest these materials were often used to improve drinking water systems). It is therefore unsurprising that the program had no significant impact on irrigation capacity. Previous remote-sensing analysis of forest area showed no landscape-scale impact of Watershared on deforestation (Wiik et al. 2019). However, our results suggest the intervention is having an impact on some relevant intermediate outcomes. The program's theory of change may thus be correct, but it is perhaps still too early to detect ultimate impacts.

Watershared aims to improve water quality by encouraging people to keep cattle out of rivers by providing barbed wire and materials to build cattle drinking troughs. Although there was no evidence of the intervention increasing the number of water intakes protected from cattle or cattle drinking points being separated from rivers ($p < 0.05$), treated and treatment households were more likely to use barbed wire to protect water intakes and to have done this more recently, suggesting that more intakes may have been protected at baseline in control communities. (We lacked data on this; however, it would be surprising to invest in protecting intakes already protected.) Regardless of a potential baseline imbalance, it is clear that water-quality-related outcomes did not materialize. The lower membership of water committees in treatment than control groups may have been because households perceived problems with water quality had been dealt with by the intervention (but this deserves further investigation).

It is interesting that the as-treated and as-randomized results were quite similar. This suggests that identified effects of Watershared were felt by the wider population in treated communities and not just those who entered agreements. This is not surprising given that several outcomes either related to outcomes independent of individual actors (e.g., perceptions of the wider environment) or related to shared resources (e.g., water intakes).

One of the key challenges in conservation impact evaluations is dealing with spillover (Baylis et al. 2016). When benefits of a program flow from treatment to control communities (through biophysical or social processes), the measured difference in outcomes of interest between the groups is reduced, making impacts of the intervention harder to detect. Accepting the risk of such spillover is inherent to any study such as ours, which treats communities within a continuous social–ecological system as randomization units. However, because spillover makes it harder to detect an intervention effect, we believe our identification of significant effects is conservative.

Overall, we found that the Watershared intervention changed land-use practices and environmental perceptions. Following the theory of change, it seems plausible that some ultimate outcomes may yet materialize. However, the impact of the intervention would likely have been enhanced with spatial targeting (Pynegar et al. 2018), increased technical support, and higher additionality (Bottazzi et al. 2018).

Given the importance of improving the effectiveness of conservation interventions, especially those that aim to deliver better social outcomes alongside environmental benefits (Sims & Alix-Garcia 2017), more robust evaluations are sorely needed (Snilsveit et al. 2019). Although RCTs are not practical or desirable in every situation and have well-understood limitations (Deaton & Cartwright

2018), our results show that the criticism that RCTs are inherently reductionist and cannot give insights into mechanisms is unjustified. By using the Watershared RCT to explore outcomes from across the intervention's theory of change, we have provided understanding of what is, and is not, changing on the ground because of the intervention. Such an analysis is inevitably complex. Preregistration (ideally alongside a peer review commitment to publish whether the results are positive or negative) is particularly important in such circumstances. We hope that preregistration becomes the norm in conservation science in the same way as is happening in other applied disciplines.

Acknowledgments

We thank colleagues at Fundación Natura Bolivia especially M. T. Vargas and T. Vidaurre. We are also grateful to K. Jack, who designed the initial randomization in 2010. This research was funded by grant RPG 2014 056 from the Leverhulme Trust and grants NE/I00436X/1 and NE/L001470/1 from the U.K.'s Ecosystem Services and Poverty Alleviation program. We thank H. Fraser and 2 anonymous reviewers who gave very helpful comments on the stage-1 manuscript. We acknowledge the support of the Supercomputing Wales project, which is partly funded by the European Regional Development Fund (ERDF) via Welsh Government.

Supporting Information

Household survey coverage (Appendix S1), outcome selection rationale (Appendix S2), definition of treated households (Appendix S3), matching variable selection rationale (Appendix S4), propensity score construction and selection (Appendix S5), matching protocol (Appendix S6), multiple testing adjustment (Appendix S7), outcome regression details (Appendix S8), outcome regression supplementary results (Appendix S9), and supplementary literature (Appendix S10) are available online. The authors are solely responsible for the content and functionality of these materials. Queries should be directed to the corresponding author. The preregistered protocol is available from https://osf.io/evymn/

Literature Cited

Asquith N, Vargas MT. 2007. Fair deals for watershed services in Bolivia. International Institute of Environment and Development, London.

Asquith NM. 2016. Watershared: adaptation, mitigation, watershed protection and economic development in Latin America. Climate & Development Knowledge Network, Cape Town, South Africa.

Baylis K, Honey-Rosés J, Börner J, Corbera E, Ezzine-de-Blas D, Ferraro PJ, Lapeyre R, Persson UM, Pfaff A, Wunder S. 2016. Mainstreaming impact evaluation in nature Conservation. Conservation Letters **9:**58–64.

Bonell C, Fletcher A, Morton M, Lorenc T, Moore L. 2012. Realist randomised controlled trials: a new approach to evaluating complex public health interventions. Social Science & Medicine **75:**2299–2306.

Börner J, Baylis K, Corbera E, Ezzine-de-Blas D, Ferraro PJ, Honey-Rosés J, Lapeyre R, Persson UM, Wunder S. 2016. Emerging evidence on the effectiveness of tropical forest conservation. PLOS ONE **11:**e0159152.

Börner J, Baylis K, Corbera E, Ezzine-de-Blas D, Honey-Rosés J, Persson UM, Wunder S. 2017. The effectiveness of payments for environmental services. World Development **96:**359–374.

Bottazzi P et al. 2017. Baseline and endline socio-economic data from a randomised control trial of the Watershared intervention in the Bolivian Andes. ReShare UK Data Archive, Colchester, United Kingdom.

Bottazzi P, Wiik E, Crespo D, Jones JPG. 2018. Payment for environmental "self-service": exploring the links between farmers' motivation and additionality in a conservation incentive programme in the Bolivian Andes. Ecological Economics **150:**11–23.

Daw JR, Hatfield LA. 2018. Matching and regression to the mean in difference-in-differences analysis. Health Services Research **53:**4138–4156.

Daw TM. et al. 2016. Elasticity in ecosystem services: exploring the variable relationship between ecosystems and human well-being. Ecology and Society **21:**art11.

Deaton A, Cartwright N. 2018. Understanding and misunderstanding randomized controlled trials. Social Science and Medicine **210:**2–21.

Ezzine-de-Blas D, Wunder S, Ruiz-Pérez M, Moreno-Sanchez R del P, Nikolakis W, Wilson P. 2016. Global patterns in the implementation of payments for environmental services. PLOS ONE **11:** e0149847.

Ferraro PJ. 2009. Counterfactual thinking and impact evaluation in environmental policy. New Directions for Evaluation **2009:**75–84.

Ferraro PJ, Hanauer MM. 2014. Advances in Measuring the environmental and social impacts of environmental programs. Annual Review of Environment and Resources **39:**495–517.

Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F. 2018. Questionable research practices in ecology and evolution. PLOS ONE **13:**e0200303 https://doi.org/10.1371/journal.pone.0200303.

Freedman LP, Cockburn IM, Simcoe TS. 2015. The economics of reproducibility in preclinical research. PLOS Biology **13:**e1002165 https://doi.org/10.1371/journal.pbio.1002165.

Glennerster R, Takavarasha K. 2013. Running randomized evaluations: a practical guide. Princeton University Press, Princeton, New Jersey.

Greifer N. 2020. cobalt: covariate balance tables and plots. R package. Version 4.0.0. https://CRAN.R-project.org/package=cobalt.

Grillos T. 2017. Economic vs non-material incentives for participation in an in-kind payments for ecosystem services program in Bolivia. Ecological Economics **131:** 178–190.

Grillos T, Bottazzi P, Crespo D, Asquith N, Jones JPG. 2019. In-kind conservation payments crowd in environmental values and increase support for government intervention: a randomized trial in Bolivia. Ecological Economics **166:**106404.

Hill J. 2008. Discussion of research using propensity-score matching: comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, *Statistics in Medicine*. Statistics in Medicine **27:**2055–2061.

Ho DE, Imai K, King G, Stuart EA, Ho DE, Imai K, King G, Stuart EA. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis **15:**199–236.

Jack BK, Kousky C, Sims KRE. 2008. Designing payments for ecosystem services: lessons from previous experience with incentive-based mechanisms. Proceedings of the National Academy of Sciences **105:**9465–9470.

Jayachandran S, de Laat J, Lambin EF, Stanton CY, Audy R, Thomas NE. 2017. Cash for carbon: a randomized trial of payments for ecosystem services to reduce deforestation. Science **357:**267–273.

Liu Z, Kontoleon A. 2018. Meta-analysis of livelihood impacts of payments for environmental services programmes in developing countries. Ecological Economics **149:**48–61.

Miteva DA, Pattanayak SK, Ferraro PJ. 2012. Evaluation of biodiversity policy instruments: what works and what doesn't? Oxford Review of Economic Policy **28:**69–92.

Open Science Collaboration OS. 2015. Estimating the reproducibility of psychological science. Science **349:**aac4716.

Parker T, Fraser H, Nakagawa S. 2019. Making conservation science more reliable with preregistration and registered reports. Conservation Biology **33:**747–750. https://doi.org/10.1111/cobi.13342.

Pynegar E. 2018. The use of randomised control trials in evaluating conservation interventions: the case of Watershared in the Bolivian Andes. Bangor University, Bangor, United Kingdom.

Pynegar EL, Gibbons JM, Asquith NM, Jones JPG. 2019. What role should randomised control trials play in providing the evidence base underpinning conservation? Oryx **6**. https://doi.org/10.1017/S0030605319000188.

Pynegar EL, Jones JPG, Gibbons JM, Asquith NM. 2018. The effectiveness of payments for ecosystem services at delivering improvements in water quality: lessons for experiments at the landscape scale. PeerJ **6:**e5753.

Samii C, Lisiecki M, Kulkarni P, Paler L, Chavis L. 2014. Effects of payment for environmental services (PES) on deforestation and poverty in low and middle income countries: a systematic review. Campbell Systematic Reviews **10:**1–95.

Sekhon JS. 2011. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. Journal of Statistical Software **42:**1–52.

Sims KRE, Alix-Garcia JM. 2017. Parks versus PES: evaluating direct and incentive-based land conservation in Mexico. Journal of Environmental Economics and Management **86:**8–28.

Snilsveit B, Stevenson J, Langer L, Da N, Zafeer S, Promise R, Polanin NJ, Shemilt I, Eyers J, Ferraro PJ. 2019. Systematic review 44. Incentives for climate mitigation in the land use sector-the effects of payment for environmental services (PES) on environmental and socio-economic outcomes in low-and middle-income countries. A mixed-method systematic review. International Initiative for Impact Evaluation. https://doi.org/10.23846/SR00044.

Streiner DL. 2015. Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests. The American Journal of Clinical Nutrition **102:** 721–728.

Wan F. 2019. Matched or unmatched analyses with propensity-score–matched data? Statistics in Medicine **38:**289–300.

White H. 2009. Theory-based impact evaluation: principles and practice. Journal of Development Effectiveness **1:**271–284.

Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York.

Wiik E, d'Annunzio R, Pynegar E, Crespo D, Asquith N, Jones JPG. 2019. Experimental evaluation of the impact of a payment for environmental services program on deforestation. Conservation Science and Practice **1:**e8.

Wilebore B, Voors M, Bulte E, Coomes D, Kontoleon A. 2019. Unconditional transfers and tropical forest conservation. Evidence from a randomized control trial in Sierra Leone. American Journal of Agricultural Economics **101:**894–918.

Wood SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73:**3–36.

Wood SN. 2017. Generalized additive models: an introduction with R. 2nd edition. CRC Press, Boca Raton, Florida.